Final Report
ANDS (Advanced Networking and Distributed Systems)
Grant MDA 972-91-J-1011
6/1/91 - 5/31/97

1. INTRODUCTION

Statement of Work:

The statement of work that describes the scope of effort the the ANDS grant is as follows:

TOPIC A: High Speed Networking

Task A1: Fast Packet Switching Using Multistage Interconnection Networks

We propose to investigate the performance of a variety of Multistage Interconnection Networks such as the Starlite network. We will develop analytical models to evaluate the throughput and response time of the overall traffic in the case of uniform traffic as well as certain forms of hot spot traffic. We will also evaluate the behavior of Message Combining to eliminate the effects of hot spots. A transformation and superposition method is being developed to be used with the analytical model to evaluate any given general traffic pattern (e.g., multiple hot spots). A delay model analysis comparing the discarding switch and the blocking switch will also be developed. We also propose to study a structured buffered pool scheme to prevent normal traffic from being blocked by the saturated tree caused by hot spot traffic.

Task A2: Analysis of Competing Lightwave Networks

The use of Wavelength Division Multiple Access (WDMA) optical switching for high-speed packet networks is a predictable development in the evolution of fast packet switching. We propose to evaluate the behavior of single-hop WDMA optical switching, using agile receiver filters. Whereas our main thrust will be on these single-hop structures, we will also look at multi-hop access using fixed filters. We will compare the response time, blocking and throughput for each.

TOPIC B: ARCHITECTURE AND PARALLEL PROCESSING

Task B1: Performance of Boolean n-Cube Interconnection Networks

We propose to evaluate the performance of Boolean n-cube interconnection networks for parallel processing systems. The focus will be on data

communication issues rather than on processing issues. By exploiting the homogeneity property of Boolean n-cube interconnection networks, we can design non-blocking routing algorithms with limited size buffers. A technique called referral is used to guarantee that every node accepts all the messages transmitted from its neighbors. This type of routing algorithm is critical in any implementation. Store-and-forward is one such routing algorithm. In this scheme, time is divided into cycles to which the network is synchronized. In each cycle every node simultaneously transmits some of its stored messages to its neighbors. An analytical model will be developed to predict the network performance under different traffic patterns. We also intend to design an intelligent routing algorithm to improve the performance. Another routing scheme to consider is a modified version of virtual cut-through. Virtual cut-through is a scheme such that when a message arrives at an intermediate node and its selected outgoing channel is free, then the message is sent to the adjacent node before it is completely received at this intermediate node. Therefore, the delay due to unnecessary buffering in front of an idle channel is avoided. Modified virtual cut-through is also a non-blocking algorithm. We will investigate the (positive or negative) effect of adding additional buffers to a node in this case. We are further interested in non-uniform traffic problems in Boolean n-cube networks.

We also propose to study the performance of these networks in a hostile and/or unreliable environment. In this environment, nodes and links may disappear and also unreliable (i.e., noisy) transmissions may occur.

Task B2: Distributed Simulation

Parallel asynchronous simulation methods (such as Time Warp) offers an optimistic alternative to synchronous conservative approaches to distributed simulation. We propose to evaluate the speedup of P processors conducting a parallel asynchronous simulation using analytic and simulation tools. We already have an exact solution for the case of two processors (P=2). Also, we have upper bounds on the best one can do by letting the P processors run ahead of each other as compared to forcing them to synchronize at every step. We are interested in extending the results to P processors and to include the effect of queued messages. Furthermore, we propose to investigate the use of the linear Poisson process as a model for these systems.

Task B3: A New Model of Load Sharing

We are interested in studying the behavior of interacting processes which gobble up processing resources in their neighborhood. In particular, if we begin with a one-dimensional world, we can place processes on a ring, where there is a quantity of processing power distributed uniformly around the ring A process requires a changing among of processing capacity. As its needs increase, the process attempts to grow in both directions along the ring until it

either has enough capacity, or it bumps into another process moving in its direction, in which case they both stop moving toward each other. As time progresses, a process may or may not have all the capacity it needs. The object is to study the response time of jobs represented by such processes in a limited resource, competitive environment. Clearly, this model extends to higher dimensions, and we propose to study the case where processors are distributed over a multi-dimensional hypersphere. The effect of distributed load sharing in this environment will be evaluated.

2. OBJECTIVE:

This objective of this contract was to investigate the architecture, design, control, load sharing and fundamental behavior of advanced networks and distributed systems from an analytical as well as from a simulation point of view, the emphasis being on analytical. In addition, algorithms and tools have been developed and implemented for these systems. The specific systems of interest include: high speed networks; parallel processing systems; and multiprocessor
systems.

3. APPROACH:

The approach was to develop mathematical models of these systems, and then to solve these models analytically or by use of simulation and or numerical methods otherwise.

The use of analytic modeling is extremely effective since it is the fastest and most economical way to investigate the effect of design decisions and alternative architectures. When such models are imbedded in operational systems, it also provides an on-line tool for systems management, control and change management. This approach has proven its effectiveness in the past for time-sharing systems, packet-switching networks, satellite networks, and packet radio networks.

This contract made considerable progress on the tasks originally proposed and, in addition, some new tasks which represent extensions, were initiated, these tasks lying on the frontier of advanced networks and distributed systems. Below, we describe the accomplishments achieved for each of these tasks.

4. SUMMARY OF WORK ACCOMPLISHED

4.1 Topic A: High Speed Networking

One of the general results achieved in this area was to develop and evaluate the critical latency/bandwidth tradeoff for gigabit networks which identified when high-speed networks are necessary for military and commercial

applications [1]. This tradeoff identified the circumstances under which gigabit networks are needed for applications which are identified only by the file length being transmitted by the application and the distance over which that file must be transmitted.

The motivation for this work is that gigabit networks force us to deal with the propagation delay due to the finite speed of light. At T1 speeds (1.5 megabits per sec), the propagation delay across the USA is forty times SMALLER than the time required to transmit a 1 megabit file. However, at a gigabit per second, the situation is completely reversed and the propagation delay is 15 times LARGER than the time to transmit the 1 megabit file!

Thus it is imperative that one must re-think a number of issues in the gigabit world; for example, the user must pay attention to his file sizes and how latency will affect his applications. The move from megabits to gigabits has taken us into a new domain which must be understood.

A major contribution of this work is to provide the understanding of this new domain as follows. A sharp boundary has been defined which separates the case in which an application is "bandwidth limited" (in which case, the application will benefit from a faster network) and the case in which it is "latency limited" (in which case, more bandwidth does not help).

This latency/bandwidth tradeoff has been widely received with much enthusiasm. We also extended this work to include the effect of latency in multi-user systems, and we have extended Amdahl's law to show the correspondence between serial work and latency in the parallel processor environment

Task A1: Fast Packet Switching Using Multistage Interconnection Networks

One accomplishment in this task was to complete the performance analysis of finite-buffered multistage interconnection switching networks such as those intended for use in emerging broadband and gigabit networks [2] [3]. We succeeded in analyzing the effect of finite memory in the switch for arbitrary traffic patterns, such as hot-spot patterns.

The famous Head-of-the-Line (HOL) problem in multistage switching networks dramatically reduces the efficiency of these networks. It is caused when a packet at the head of an input queue is blocked because some packet on another input queue is currently being transmitted through the switching fabric to the same output queue to which the first packet is destined. The HOL problem is the phenomenon that the second (or deeper) packet in the first queue is destined for an output queue that is free, but this packet cannot be transmitted because the first packet is HOL-blocked. A novel architecture to relieve the HOL problem, known as the odd-even queue was studied in this

contract. It creates two queues at each input port, the first containing packets destined for odd numbered output ports, and the second containing the even ones. Significant gains can be achieved using this and generalized architectures of this type [4] [5].[6].

Task A2: Analysis of Competing Lightwave Networks

A major result in this task succeeded in establishing the fundamental limits on the performance of a perfect optical switch, these limits arising from constraints due to the number of tunable and fixed transmitters and receivers at the input and output of the switch [7],[8].

In addition, we developed and analyzed a novel access protocol for high-speed optical local area networks; these networks are likely to appear in the coming broadband environment of gigabit and terabit per second networks.
A new protocol for passive star optical networks was analyzed that has application to high speed local area networks [9], [10].

A multi-channel protocol for use in optical Wave Length Division Multiplexing Access (WDMA) networks has been developed which allows one to exploit the enormous bandwidth available in fiberoptic networks [11]. In an attempt to achieve these bandwidths, one is typically defeated by the bottleneck when attempting to achieve bandwidths commensurate with the potential of the full optical bandwidth due to the limitations of electronic speeds. Through the use of multiple channels (based on WDMA protocols), one is able to overcome this bottleneck and achieve much higher speeds. The multi-channel protocol developed is adapted to the emerging popular distributed queue dual bus (DQDB) protocol which has been developed for metropolitan area networks (MANs). Indeed, this protocol allows one to run a linear optical fiber bus to connect user stations in a large geographical region. The protocol was specified and an analysis was carried out which provides excellent approximate results for the throughput and delay of this multi-channel protocol. One is able to achieve throughput gains over the single channel case which are proportional to the number of wavelengths used. Moreover, this multi-channel protocol exhibits a higher degree of fairness among the attached stations than does single-channel DQDB (fairness is one of the drawbacks to single-channel DQDB).

4.2 Topic B: Architecture and Parallel Processing

The performance of parallel processing systems remains far below the potential gains offered by these systems. In order to understand and remove some of their limitations, we studied a number of architectures in this contract. In one case, developed a model for parallelism which totally generalized Amdahl's model and allowed far more generality to be analyzed [12]. In this

study, we were able to identify optimal design points for these systems. A more difficult model was developed for which we were able to obtain approximate results which identified the performance gains to be had by these systems [13].

In another study, we developed and analyzed a model of loosely coupled networks of work stations (NOW) and were able to show the gains to be had by recapturing the unused capacity of these stations when they are not being used by those at the workstation terminals themselves [14].

Task B1: Performance of Boolean n-Cube Interconnection Networks

In this contract, we developed an algorithm and evaluated the performance of a deadlock-free routing algorithm for hypercube interconnection networks with finite buffers [15]. We also developed and evaluated fault-tolerance properties of such algorithms [16]; such behavior is necessary in military applications.

Developed an algorithm and evaluated the performance of a deadlock-free routing algorithm for hypercube interconnection networks with finite buffers. Also developed and evaluated fault-tolerance for such algorithms as are necessary in military applications.

Task B2: Distributed Simulation

The time-warp algorithm is one of the most popular implementations of optimistic simulation. Its performance could not be evaluated analytically until the work we published wherein we provided the first exact analysis for two-processor time-warp simulation [17]. Further work conducted on this contract provided a number of generalizations [18], [19], [20].[21]. This work has proven itself valuable as an aid in comparing alternative simulation options.

A new approach to the general problem of robust and fault-tolerant distributed control was developed in this contract effort [22], [22]. The application here is to distributed communications as well as to distributed assignment of tasks. These algorithms function in hostile and/or non-centralized environments such as for autonomous vehicles. This control procedure allows a collection of very loosely coupled processes to accomplish a common goal effectively without much interaction among the processes. The underlying mechanism for each process is a finite state machine embedded in a broadcast communication medium. In this effort, we developed, analyzed and simulated the behavior of a highly robust distributed algorithm, namely the Goore Game. Analysis was carried out and simulation was used to corroborate the analysis. A number of novel tasks were shown to be compatible with these distributed automata including applications communications and the stabilization of accelerating objects of various shaped configurations, among others. The control algorithm responded to dynamics very well.

In this contract, we also unified and extended the work on Petri nets for the performance evaluation of distributed systems. Emphasis here was on the development of efficient algorithms for structural level specification of these nets, and to develop bounds based on these structural constructs of the model. The goal was to overcome the current computational bottleneck in the application of Petri nets to high-performance processors and communications [24].

Task B3: A New Model of Load Sharing

A new model of load sharing was developed and analyzed. We also established analytical and simulation results for the gains to be had by using a new parallel systems architecture, namely a Virtual Time Data-Parallel machine [25] in which parallelism is explicitly modeled

Access to shared databases by a number of competing users was modeled and analyzed using a variety of algorithms. This led to the notion of Winner Queues [26].

5. CONCLUSION

The research conducted on this contract produced a wealth of results which have had impact on the computer and communications technology base.

The number of professional publications in refereed journals was 25. The number of Ph.D. students supported was nine.

Some fundamental results were developed which identify the ultimate limits and tradeoffs to be considered when designing these systems. The areas of investigation covered all of the contracted tasks in the statement of work, and went beyond these to uncover further principles and understanding.

6. REFERENCES

1. Kleinrock, L., "The Latency/Bandwidth Tradeoff In Gigabit Networks", *IEEE Communications Magazine*, April 1992, Vol. 30, No. 4, pp.36-40.

2. Lin, T.I. and L. Kleinrock, "Performance Analysis of Finite-Buffered Multistage Interconnection Networks with A General Traffic Pattern," *1991 ACM Sigmetrics, Conference on Measurement and Modeling of Computer Systems*, Vol. 19, No. 1, pp. 68-78, May 21-24, 1991, San Diego, CA.

3. Lin, T-I., and Kleinrock, L., "Performance Analysis of the Finite-Buffered 'Turn-Back' Multistage Interconnection Network", *IFIP Workshop TC6*, La Martinique, French Caribbean Island, pp. 3-22, January 25-27, 1993.

4. Kolias, C., L. Kleinrock, "The Odd-Even Input-Queueing ATM Switch: Performance Evaluation", *ICC'96*, June 23-27, pp. 1674-1679, 1996.

5. Kolias, C., and L. Kleinrock, "The Dual-Banyan (DB) Switch: A High Performance Buffered-Banyan ATM Switch", *ICC'97*, June 1997, Montreal, Canada, pp. 770-776.

6. Kolias, C., L. Kleinrock, "Throughput Analysis of Multiple Input-Queueing in ATM Switches", *IFIP-IEEE Broadband Communications*, Eds, L. Mason and A. Casaca, pp. 382-393, Chapman and Hall, 1996.

7. Lu, J. and L. Kleinrock, "On The Performance Of Wavelength Division Multiple Access Networks", *ICC'92*, Chicago, IL, pp. 1151-1157.

8. Lu, J.C. and L. Kleinrock "Performance Analysis of Single-Hop Wavelength Division Multiple Access Networks" *Journal of High-Speed Networks*, Vol. 1, No. 1, 1992, pp.61-77, 1992.

9. Lu, Jonathan and L. Kleinrock, "An Access Protocol For High-Speed Optical LANs", *Proceedings of the 20th Annual Computer Science Conference*, March 3-5, 1992, Kansas City, Missouri, pp. 287-293.

10. Lu, J. and L. Kleinrock, "A Wavelength Division Multiple Access Protocol for High-speed Local Area Networks with a Passive Star Topology," *Performance Evaluation*, Vol. 16, No.1-3, pp. 223-239, November 1992.

11. Lu, J. C. and Kleinrock, L., "A WDMA Protocol for Multichannel DQDB Networks" *GLOBECOM '93*, pp. 149-153, January 1993.

12. Kleinrock, L. and J. H. Huang, "On Parallel Processing Systems: Amdahl's Law Generalized and Some Results on Optimal Design", invited paper for *IEEE Transactions on Software Engineering*, Special issue on Performance Evaluation Methodology. Vol. 18, No. 5, May 1992, pp. 434-447.

13. Huang, J.H. and L. Kleinrock, "Performance Evaluation of Dynamic Sharing of Processors in Two-Stage Parallel Processing Systems," *IEEE Transactions on Parallel and Distributed Systems*, Vol. 4, No. 3, March 1993, pp. 306-317.

14. Kleinrock, L. and W. Korfhage, "Collecting Unused Processing Capacity: An Analysis of Transient Distributed Systems", *IEEE Transactions on Parallel and Distributed Systems*, May 1993, pp. 535-546.

15. Horng, Ming-yun, and L. Kleinrock, "On the Performance of a Deadlock-free Routing Algorithm for Boolean n-Cube Interconnection Networks with Finite Buffers," *1991 International Conference on Parallel Processing*, Austin, TX, pp. III-228-III-232, August 12-16, 1991.

16. Horng, Ming-yun and L. Kleinrock, "Fault-Tolerant Routing With Regularity Restoration In Boolean n-Cube Interconnection Networks", *Proceedings of the Third IEEE Symposium on Parallel and Distributed Processing*, Dallas, Texas, December 2-5, 1991, pp. 458-465.

17. Kleinrock, L. "On Distributed Systems Performance", presented at the 7th ITC Specialist Seminar in Australia, *Proceedings of the ITC Specialist Seminar*, September 1989.

18. Felderman, R. and L. Kleinrock, "Two Processor Conservative Simulation Analysis," *1992 PADS Workshop*, Newport Beach, CA, June 1991. (Also, Information Science Institute Tech. Rept. ISI/RS-92-299).

19. Felderman, R. and L. Kleinrock, "Bounds and Approximations for Self-Initiating Distributed Simulation Without Lookahead", *ACM Transactions on Modelling and Computer Simulation*, special issue on Distributed and Parallel Simulation Performance, Vol. 1, No.4., 1991, pp. 386-406. (Also, Information Sciences Institute Tech. Rept. ISI/RS-92-298).

20. Kleinrock, L. and R. Felderman, "Two Processor Time Warp Analysis: A Unifying Approach," *International Journal in Computer Simulation*, Volume 2, Number 4, pp. 345-371, 1992.

21. Felderman, R. and L. Kleinrock, "Two Processor Time Warp Analysis: Capturing the Effects of Message Queueing and Rollback/State Saving Costs," Memorial Issue for Felix Pollaczek. AEU special issue *"Teletraffic Theory and Engineering"*, September/November 1993, Vol. 47, Issue 5-6, pp. 353-367.

22. Tung, B. and L. Kleinrock, "Distributed Control Methods*", 2nd International Symposium on High Performance Distributed Computing*, Spokane, Washington, July 21-23, 1993, pp. 206-215.

23. Tung, B. and L. Kleinrock "Using Finite State Automata to Produce Self-Optimization and Self Control", *IEEE Transactions on Parallel and Distributed Systems*, Vol. 7, No. 4, pp. 439-448, April 1996.

24. Kleinrock, L. and D. Nielsen, "Data Structures and Algorithms for Extended State Space and Structural Level Reduction of the GSPN Model", *15th International Conference on Application and Theory of Petri Nets*, Zaragoza, June 20-24, 1994.

25. Shen, S. and L. Kleinrock, "The Virtual Time Data Parallel Machine", *4th Symposium, on the Frontiers of Massively Parallel Computation*, 1992, IEEE Computer Society, McLean, VA, 1992, pp. 46-53.

26. Kleinrock, L. and F. Mehovic, "Poisson Winner Queues," *Performance Evaluation*, Vol. 14, No. 2, 1992, pp. 79-101.